

1 1. A hierarchical clustered parallel processing system comprising:

2 at least one cluster of computer processing system forming a node of a
3 hierarchical cluster, each cluster of computer processing systems
4 comprising:

5 a plurality of computer systems designated to be members of nodes
6 of said cluster;

7 a physical network connected to allow each computer system of the
8 plurality of computer systems to transfer data between any of
9 the plurality of computer systems;

10 a virtual multicast bus to designate communication between
11 member computer systems; and

12 a configuration service apparatus in communication with each of the
13 computer systems to provide each of the plurality of computer
14 systems with:

15 a node identification to identify a node for each member
16 computer system within said cluster,

17 a multicast bus address to broadcast communications to
18 said members of said cluster by way of said virtual
19 cluster bus, and

20 a node priority list designating a priority for each node within
21 said cluster; and

22 a cluster supervising processor to provide operational control
23 services for said cluster, said cluster supervising processor
24 being selected of said member computer systems according to
25 the priority from said priority list.

1 2. The hierarchical clustered parallel processing system of claim 1 wherein the
2 configuration service apparatus further provides a disk access list.

1 3. The configuration service apparatus of claim 2 wherein the disk access list
2 comprises identification of accessible disks, disk mount points, and failure
3 detection locations.

1 4. The hierarchical clustered parallel processing system of claim 1 wherein the
2 cluster supervising processor maintains:

3 a cluster topology table detailing connectivity for each node of the cluster
4 and a disk access status for each disk within said cluster;

5 a disk usage table describing current capacity and loading for each disk
6 within said cluster;

7 a node usage table describing a streaming capacity for each node of said
8 cluster and a current loading for each node of said cluster; and

9 a cluster map describing network addresses for each of a plurality of
10 servers in communication with said cluster and listing of nodes within
11 said cluster, network addresses for said nodes, and an operational
12 status of said nodes.

1 5. The hierarchical clustered parallel processing system of claim 1 wherein a group
2 of said member computer systems of said cluster are configured as a sub-
3 cluster, said sub-cluster being a node of said cluster.

1 6. The hierarchical clustered parallel processing system of claim 4 wherein the each
2 cluster of computer processing systems further comprises a fault detection
3 apparatus within each member computer system:

4 to periodically receive a first processor status message from a first
5 adjacent node;

6 to append a second processor status message of a current node to said
7 first processor status message; and

8 to periodically transmit said first and second processor status message to
9 a second adjacent node.

1 7. The hierarchical clustered parallel processing system of claim 6 wherein said
2 cluster supervising processor receives an accumulation of the processor status
3 messages from all nodes of said cluster.

- 1 8. The hierarchical clustered parallel processing system of claim 6 wherein, if the
2 fault detection apparatus does not receive said first processor status message for
3 a number of periods, said first adjacent node is declared to have failed and a
4 failure declaration is appended to said second processor status message.
- 1 9. The hierarchical clustered parallel processing system of claim 8 wherein, upon
2 receipt of said failure declaration, the cluster supervising processor modifies said
3 cluster map to reflect failure of the node.
- 1 10. The hierarchical clustered parallel processing system of claim 4 wherein the
2 cluster supervising processor periodically posts a supervisor notification
3 message on said virtual multicast bus, said supervisor notification message
4 comprises a node identification and a network address for said cluster
5 supervising processor.
- 1 11. The hierarchical clustered parallel processing system of claim 10 wherein the
2 supervisor notification message further comprises the cluster topology and a
3 current cluster map.
- 1 12. The hierarchical clustered parallel processing system of claim 10 wherein, if one
2 node of cluster does not receive said supervisor notification message within a
3 notification time, said node becomes said cluster supervising processor, updates
4 said cluster topology table and said cluster map, transmits a cluster supervising
5 processor update message, and the supervisor notification message.
- 1 13. The hierarchical clustered parallel processing system of claim 4 wherein:

2 each node of said cluster periodically determines whether each disk to
3 which said node has access is functioning and if any disk is not
4 functioning;

5 the node creates a disk failure message for the disk not functioning for
6 transfer to an adjacent node;

7 wherein said adjacent node transfers said disk failure node to subsequent
8 adjacent nodes until said cluster supervising processor receives said
9 disk failure message;

10 wherein upon receipt of multiple disk failure messages from multiple
11 nodes for the disk not functioning, the cluster supervising processor
12 declares a disk failure, updates the disk usage table, and reassigns all
13 the transfer of video data files from a failing node to an active node.

1 14. The hierarchical clustered parallel processing system of claim 10 wherein a new
2 node joins said cluster by the steps of:

3 listening to said virtual multicast bus for a supervisor notification message
4 from the present cluster supervising processor;

5 posting on said virtual multicast bus a join request message providing a
6 node identification, a network address for said node, and a disk access
7 list for said node;

8 updating by the present cluster supervising processor the cluster map and
9 the cluster topology; and
10 placing a new supervisor notification message upon said virtual multicast
11 bus including said new node.

1 15. The hierarchical clustered parallel processing system of claim 14 wherein the
2 new node joins said cluster further by the step of:

3 ceasing posting on said virtual multicast bus said join request message.

4 16. The hierarchical clustered parallel processing system of claim 14 wherein the
5 new node becomes the cluster supervising processor, if said new node has a
6 priority that supercedes said present cluster supervising processor.

1 17. The hierarchical clustered parallel processing system of claim 16 wherein the
2 new node acting as the present cluster supervising processor transmits the
3 supervisor notification message and the original cluster supervising processor
4 ceases transmitting said supervisor notification message.

1 18. The hierarchical clustered parallel processing system of claim 17 wherein if the
2 new node does not transmit the supervisor notification message by the
3 notification time, the original cluster supervising processor resumes transmission
4 of the supervisor notification message.

1 19. The hierarchical clustered parallel processing system of claim 10 wherein a node
2 leaves a cluster by the steps of:

3 posting a leave message on said virtual multicast bus, said leave
4 message containing the node identification and the network address
5 for said node;

6 updating by the cluster supervising processor of the cluster map and the
7 cluster topology; and

8 posting on the virtual multicast bus the supervisor notification message
9 with the updated cluster map and cluster topology.

1 20. The hierarchical clustered parallel processing system of claim 19 wherein the
2 node leaving the cluster ceases posting the leave message upon receipt of the
3 supervisor notification message with the updated cluster map and cluster
4 topology.

1 21. The hierarchical clustered parallel processing system of claim 19 wherein if the
2 node leaving the cluster is the cluster supervising processor, the node of the
3 cluster of the priority list then becomes the cluster supervising processor.

1 22. The hierarchical clustered parallel processing system of claim 1 wherein said
2 cluster is formed and said cluster supervising processor is designated by the
3 steps of:

4 listening to said virtual multicast bus for a supervisor notification message
5 from the cluster supervising processor by each node of the cluster;

6 if no supervisor notification message is received, designating each node a
7 single node cluster of its own;

8 designating each node the cluster supervising processor of its single node
9 cluster;

10 transmitting by each cluster supervising processor of each single node
11 cluster the supervisor notification message for each single node
12 cluster;

13 ceasing by those nodes having a lower priority from transmitting
14 supervisor notification messages such that the node with a highest
15 priority is the cluster supervising processor; and

16 joining said cluster by those nodes with lower priority by posting on said
17 virtual multicast bus a join request message providing a node
18 identification, a network address for said node, and a disk access list
19 for said node.

- 1 23. A cluster of computer processing systems comprising:
- 2 a plurality of computer systems designated to be members of nodes of
3 said cluster;
- 4 a physical network connected to allow each computer system of the
5 plurality of computer systems to transfer data between any of the
6 plurality of computer systems;

7 a virtual multicast bus to provide communication between member
8 computer systems; and
9 a configuration service apparatus in communication with each of the
10 computer systems to provide each of the plurality of computer systems
11 with:
12 a node identification to identify a node for each member
13 computer system within said cluster,
14 a multicast bus address to broadcast communications to said
15 members of said cluster by way of said virtual cluster bus,
16 and
17 a node priority list designating a priority for each node within
18 said cluster; and
19 a cluster supervising processor to provide operational control services for
20 said cluster, said cluster supervising processor being selected of said
21 member computer systems according to the priority from said priority
22 list.

1 24. The cluster of computer processing systems of claim 23 wherein the
2 configuration service apparatus further provides a disk access list.

1 25. The configuration service apparatus of claim 24 wherein the disk access list
2 comprises identification of accessible disks, disk mount points, and failure
3 detection locations.

1 26. The cluster of computer processing systems of claim 23 wherein the cluster
2 supervising processor maintains:

3 a cluster topology table detailing connectivity for each node of the cluster
4 and a disk access status for each disk within said cluster;

5 a disk usage table describing current capacity and loading for each disk
6 within said cluster;

7 a node usage table describing a streaming capacity for each node of said
8 cluster and a current loading for each node of said cluster; and

9 a cluster map describing network addresses for each of a plurality of
10 servers in communication with said cluster and listing of nodes within
11 said cluster, network addresses for said nodes, and an operational
12 status of said nodes.

1 27. The cluster of computer processing systems of claim 23 wherein a group of said
2 member computer systems of said cluster are configured as a sub-cluster, said
3 sub-cluster being a node of said cluster.

1 28. The cluster of computer processing systems of claim 26 further comprising a fault
2 detection apparatus within each member computer system:

3 to periodically receive a first processor status message from a first
4 adjacent node and transmit;

5 to append a second processor status message of a current node to said
6 first processor status message; and

7 to periodically transmit said first and second processor status message to
8 a second adjacent node.

1 29. The cluster of computer processing systems of claim 28 wherein said cluster
2 supervising processor receives an accumulation of the processor status
3 messages from all nodes of said cluster.

1 30. The cluster of computer processing systems of claim 28 wherein, if the fault
2 detection apparatus does not receive said first processor status message for a
3 number of periods, said first adjacent node is declared to have failed and a
4 failure declaration is appended to said second processor status message.

1 31. The cluster of computer processing systems of claim 30 wherein, upon receipt of
2 said failure declaration, the cluster supervising processor modifies said cluster
3 map to reflect failure of the node.

1 32. The cluster of computer processing systems of claim 26 wherein the cluster
2 supervising processor periodically posts a supervisor notification message on
3 said virtual multicast bus, said supervisor notification message comprises a node
4 identification and a network address for said cluster supervising processor.

1 33. The cluster of computer processing systems of claim 32 wherein the supervisor
2 notification message further comprises the cluster topology and a current cluster
3 map.

1 34. The cluster of computer processing systems of claim 32 wherein, if one node of
2 cluster does not receive said supervisor notification message within a notification
3 time, said node becomes said cluster supervising processor, updates said cluster
4 topology table and said cluster map, transmits a cluster supervising processor
5 update message, and the supervisor notification message.

1 35. The cluster of computer processing systems of claim 26 wherein:

2 each node of said cluster periodically determines whether each disk to
3 which said node has access is functioning and if any disk is not
4 functioning;

5 the node creates a disk failure message for the disk not functioning for
6 transfer to an adjacent node;

7 wherein said adjacent node transfers said disk failure node to subsequent
8 adjacent nodes until said cluster supervising processor receives said
9 disk failure message;

10 wherein upon receipt of multiple disk failure messages from multiple
11 nodes for the disk not functioning, the cluster supervising processor

12 declares a disk failure, updates the disk usage table, and reassigns all
13 the transfer of video data files from a failing node to an active node.

1 36. The cluster of computer processing systems of claim 32 wherein a new node
2 joins said cluster by the steps of:

3 listening to said virtual multicast bus for a supervisor notification message
4 from the present cluster supervising processor;

5 posting on said virtual multicast bus a join request message providing a
6 node identification, a network address for said node, and a disk access
7 list for said node;

8 updating by the present cluster supervising processor the cluster map and
9 the cluster topology; and

10 placing a new supervisor notification message upon said virtual multicast
11 bus including said new node.

12 37. The cluster of computer processing systems of claim 36 wherein the new node
13 joins said cluster further by the steps of:

14 ceasing posting on said virtual multicast bus said join request message.

15 38. The cluster of computer processing systems of claim 36 wherein the new node
16 becomes the cluster supervising processor, if said new node has a priority that
17 supercedes said present cluster supervising processor.

- 1 39. The cluster of computer processing systems of claim 38 wherein the new node
2 acting as the present cluster supervising processor transmits the supervisor
3 notification message and the original cluster supervising processor ceases
4 transmitting said supervisor notification message.
- 1 40. The cluster of computer processing systems of claim 39 wherein if the new node
2 does not transmit the supervisor notification message by the notification time, the
3 original cluster supervising processor resumes transmission of the supervisor
4 notification message.
- 1 41. The cluster of computer processing systems of claim 32 wherein a node leaves a
2 cluster by the steps of:
- 3 posting a leave message on said virtual multicast bus, said leave
4 message containing the node identification and the network address
5 for said node;
- 6 updating by the cluster supervising processor of the cluster map and the
7 cluster topology; and
- 8 posting on the virtual multicast bus the supervisor notification message
9 with the updated cluster map and cluster topology.
- 1 42. The cluster of computer processing systems of claim 41 wherein the node
2 leaving the cluster ceases posting the leave message upon receipt of the

3 supervisor notification message with the updated cluster map and cluster
4 topology.

1 43. The cluster of computer processing systems of claim 41 wherein if the node
2 leaving the cluster is the cluster supervising processor, the node of the cluster of
3 the priority list then becomes the cluster supervising processor.

1 44. The cluster of computer processing systems of claim 23 wherein said cluster is
2 formed and said cluster supervising processor is designated by the steps of:

3 listening to said virtual multicast bus for a supervisor notification message
4 from the cluster supervising processor by each node of the cluster;

5 if no supervisor notification message is received, designating each node a
6 single node cluster of its own;

7 designating each node the cluster supervising processor of its single node
8 cluster;

9 transmitting by each cluster supervising processor of each single node
10 cluster the supervisor notification message for each single node
11 cluster;

12 ceasing by those nodes having a lower priority from transmitting
13 supervisor notification messages such that the node with a highest
14 priority is the cluster supervising processor; and

15 joining said cluster by those nodes with lower priority by posting on said
16 virtual multicast bus a join request message providing a node
17 identification, a network address for said node, and a disk access list
18 for said node.

19